JLIS.it

Le edizioni digitali come nuovo modello per dati di autorità concettuali

Francesca Tomasi

1 Introduzione

La progressiva estensione degli àmbiti di intervento computazionale agli oggetti del patrimonio culturale ha determinato un'attenzione maggiore al documento inteso come dato la cui capacità espressiva va oltre la sola descrizione metadatale a livello paratestuale. La trascrizione, per esempio, sta entrando nel circuito della rappresentazione del contenuto informativo di cui libri e documenti sono portatori. Sia in campo archivistico che librario l'attenzione verso il full-text ha obbligato a tradurre il sistema di metadatazione descrittivo, amministrativo-gestionale e strutturale, che si esprime comunemente al livello del paratesto, al livello del testo. E il metadato inizia così a configurarsi come un elemento di annotazione che può trasformare il testo, sia esso documento archivistico o fonte libraria, in edizione.

L'edizione digitale di un documento può essere intesa, attraverso l'annotazione, come un processo che porta alla progressiva stratificazione del sistema interpretativo dell'editore, in modo particolare nei sistemi di markup dichiarativo (Coombs, Renear e DeRose). I

JLIS.it. Vol. 4, n. 2 (Luglio/July 2013).
DOI: 10.4403/jlis.it-8808

diversi aspetti dell'analisi dei contenuti di un documento conducono alla creazione di una raccolta di informazioni multilivellari che nascono dal processo interpretativo. Tale processo è il modello del documento, inteso come oggetto informativo complesso, elaborato dall'editore critico. Tipicamente persone, luoghi, date, oggetti, eventi e parole chiave rappresentano istanze interpretative che si configurano come elementi dell'annotazione riferiti a valori che si presentano nella forma di stringhe di caratteri. Ogni stringa interpretata o annotata (composta da elemento descrittivo e valore associato) è potenzialmente un'informazione autonoma, legata al testo dell'edizione, necessaria a fornire i diversi punti di accesso al documento ovvero a determinare le possibili entries. Tale approccio è la base di partenza per creare liste controllate di valori di elementi, estraendo dal documento sia le forme attestate che le forme varianti di nomi di persona, di luoghi, date, titoli e soggetti, per associarle quindi alla forma controllata secondo lo standard adottato. Ma ogni stringa annotata (per esempio una stringa identificabile come un "nome di persona") richiama una serie di informazioni che vanno oltre la semplice annotazione e tali informazioni provengono sia dal contesto specifico di occorrenza della stringa che da fonti esterne (per esempio luogo e data di nascita, occupazione, relazioni con altre persone). E soprattutto gli elementi annotati non solo sono in relazione fra di loro, ma intrattengono anche relazioni con altre risorse distribuite. Si passa dall'edizione digitale alla raccolta di descrizioni di dati altamente strutturati che si possono caratterizzare come un nuovo modello di authority file, in cui il punto di accesso al documento è l'esito di una relazione fra elementi annotati in un determinato contesto testuale. L'authority si trasforma così da stringa a concetto e il processo di concettualizzazione è il risultato dell'accoppiata elemento-valore e della rete di collegamenti interni (fra elementi) ed esterni (far elementi e risorse distribuite).

In prima battuta diremo quindi che gli elementi annotati andranno posti in relazione attraverso adeguati predicati ontologici. Perché una stringa identificata come ”data” e una identificata come ”persona”, o ”luogo” o ”evento” potrebbero avere una qualche connessione. Non è sufficiente un generico collegamento non tipizzato o sintattico, ma va specificata la ragione della relazione, individuando formalmente la tipologia di connessione fra gli elementi. La conoscenza che implicitamente nasce dalla lettura del documento viene così formalizzata attraverso relazioni semantiche esplicite: per esempio una data stabilisce il momento del trasferimento di una persona in un luogo; un luogo determina uno spazio in un cui un evento è stato organizzato da una persona; un soggetto identifica una feature di una persona. In secondo luogo ogni stringa annotata, oltre ad avere relazioni con altre stringhe interne al documento, ha relazioni con altri oggetti distribuiti che si riferiscono al medesimo contenuto informativo, sia a livello di singolo elemento (la stessa persona) che, soprattutto, a livello di concetto espresso in quel documento (una persona che intrattiene una relazione con un'altra persona in uno specifico contesto testuale).

Le persone, i luoghi, le date, i soggetti, gli eventi e gli oggetti vanno descritti secondo gli standard in uso, vanno messi in relazione fra di loro ad esprimere asserzioni, determinando concetti, e vanno relazionati con altre entità su WWW — che possono anche condividere lo stesso tipo di relazioni interne al documento — creando collegamenti incrociati.

Questo significa che le edizioni digitali devono confrontarsi con il mondo dei sistemi di metadazione in uso nel settore del cultural heritage, con i linguaggi formali del semantic web e con il crescente fenomeno linked data. Le edizioni digitali sono una base di conoscenza ”naturaliter” linked. Le relazioni fra le stringhe annotate nascono cioè spontaneamente, all'atto della lettura del testo. Diremo

che il contesto in cui ogni stringa occorre rappresenta le ragioni del collegamento e stabilisce il dominio di riferimento. Contesto e dominio sono due concetti chiave nella trasformazione dell'annotazione in base di conoscenza perché identificano l'àmbito di modellazione dell'edizione.

Contesto in letteratura significa che le relazioni fra stringhe nascono dall'ambito semantico in cui tali stringhe compaiono (Lee). Le relazioni che possono essere formalizzate derivano quindi dalla specifica co-occorrenza di stringhe. Ma contesto è anche un concetto che richiama inevitabilmente lo standard ISAAR-CPF[1] e la sua formalizzazione EAC-CPF.[2] Il ruolo di ISAAR-CPF in particolare diventa importante nel processo di identificazione univoca di entità come persone e come relazioni fra persone, veicolando i concetti di soggetto produttore (sia esso persona, famiglia o ente), di relazione fra il soggetto produttore e gli oggetti prodotti (vale a dire le risorse di cui il soggetto assume una forma di paternità) e di collegamento fra soggetti produttori o in generale fra persone. Di EAC-CPF peraltro c'è l'ontologia recentemente proposta che formalizza classi e proprietà dello schema (Mazzini e Ricci).[3] Ce lo insegna l'archivistica "separating description of people from description of record" (Pitti) che in campo di edizione può essere tradotto nel separare la descrizione delle persone dal testo dell'edizione, ma mantenendo il collegamento fra la persona e il documento in cui quella persona occorre, che stabilisce il contesto. Affermazione, quella di Pitti, che può essere estesa dalle persone a ogni fenomeno dell'analisi. E

[1] International Standard Archival Authority Record for Corporate Bodies, Persons and Families. Second Edition, 2003. http://www.icacds.org.uk/eng/ISAAR(CPF)2ed.pdf.

[2] Encoded Archival Context – Corporate Bodies, Persons and Families. La specifica dello schema si può leggere all'indirizzo: http://eac.staatsbibliothek-berlin.de.

[3] EAC-CPF Vocabulary Specification 1.0 si può leggere all'indirizzo: http://archivi.ibc.regione.emilia-romagna.it/ontology/reference_document/referencedocument.html.

trasformare le entità e le loro relazioni in ontologie significa trasformare i testi in basi di conoscenza. Le edizioni digitali diventano allora sistemi su cui sviluppare forme di "knowldege representation" (Clement).

Ed entriamo così nel concetto di dominio come spazio di riferimento semantico. L'ontologia è per sua stessa natura una concettualizzazione di una realtà osservata rispetto ad un àmbito di riferimento. Allo stesso modo diverse edizioni di testi se avranno fra loro entità comuni (la stessa persona, lo stesso luogo, la stessa keyword) potranno avere relazioni diverse a seconda dell'àmbito in cui queste entità compaiono. Il concetto di ontologia di dominio deve fare quindi i conti sia con la realtà osservata rispetto allo specifico contesto, sia con il punto di vista assunto sull'oggetto dell'analisi. Scopo del presente contributo è quindi di: ragionare sulle entità, nella forma di stringhe estratte da un testo annotato (elemento-valore), come entries, e quindi come punti di accesso al documento, e ragionare su come queste ultime possono configurarsi come authority files; ragionare su come estendere il concetto di authority a quello di relazione in quanto ogni authority è legata ad un contesto e ad un dominio; ragionare sul concetto di relazione come collegamento fra le authorities così configurate nello spazio del WWW in un sistema di interlinking. Tentare quindi di "andare oltre le colonne d'Ercole" (Crupi) diventa lo scopo del processo che si intende qui descrivere.

Le edizioni digitali fanno parte del patrimonio culturale e vanno quindi valorizzate al pari delle raccolte librarie, archivistiche e museali, anche in considerazione della realizzazione di digital libraries nella forma di aggregatori di risorse come strumento di accesso integrato al patrimonio culturale (come è ad esempio Europeana,[4] v. Aloia, Concordia e Meghini).Il metadato aggregato non sarà quindi più solo un elemento estratto dalla descrizione della

[4]Il portale può essere consultato all'indirizzo: http://www.europeana.eu/portal.

risorsa, ma un elemento che proviene dal testo pieno dell'oggetto digitale. Le edizioni digitali in campo letterario possono fornire ai sistemi archivistici e librari un modello già testato e oggetto di studi e sperimentazioni che può favorire il processo di trascrizione integrale delle fonti documentali e librarie. Se il processo di dialogo avviato fra archivi, biblioteche e musei[5] si estendesse al settore delle digital humanities il patrimonio culturale amplierebbe le prospettive di interesse allargando la base di conoscenza a disposizione dell'utente finale. Le già esistenti authorities in settore archivistico e librario potranno poi essere arricchite di nuovi dati provenienti da nuove fonti ancora inesplorate.

2 Il panorama di riferimento

Nel campo delle edizioni digitali di testi si registra un numero crescente di sperimentazioni (Sahle). Solo per fare qualche esempio si può esplorare la classificazione delle edizioni del XIX secolo inglesi ed americane fatta da Nines.[6] o si possono consultare i numerosi progetti editoriali del DDH (Department of Digital Humanities) del King's College di Londra;[7] si possono anche vedere i lavori del CDS (Center for Digital Scholarship) della Brown University, come lo storico Women's Writers Project,[8] o accedere ai progetti dei vari centri che si occupano di digital humanities[9] o ancora consultare l'elenco

[5]Come dimostra l'interessante progetto italiano MAB (Musei, Archivi e Biblioteche): http://www.mab-italia.org.

[6]Networked Infrastructure for Nineteenth-Century Electronic Scholarship: http://www.nines.org. Si tratta di un aggregatore di metadati provenienti da "peer-reviewed digital objects'.

[7]http://www.kcl.ac.uk/artshums/depts/ddh/research/index.aspx.

[8]http://www.wwp.brown.edu.

[9]Una classificazione si può leggere sul sito CenterNet: http://digitalhumanities.org/centernet.

di edizioni, e di progetti di digital libraries o collezioni digitali in generale, che si basano su XML/TEI[10] sullo stesso sito dedicato allo schema.[11]

Anche le istituzioni archivistiche hanno avviato procedure di trascrizione integrale delle fonti,[12] arrivando al livello dell'item come nel progetto Datini, datato 2002, condotto sulla porzione dell'omonimo fondo delle lettere di Margherita Datini a Francesco di Marco[13] o come nel lavoro sul Codice diplomatico della Lombardia medievale.[14] O ancora non si può non menzionare, in campo di trascrizione di manoscritti, l'egregio lavoro di UCL (University College London) su Jeremy Bentham[15] come esempio di progetto collaborativo in un'ottica di "social edition" (Siemens et al.).

Anche il rapporto fra le edizioni e il ruolo delle tecnologie legate al semantic web ha portato alla realizzazione di prodotti digitali di eccellenza, come, per fare un esempio, il Discovery Project (D'Iorio e Barbera)[16] relativo alla filosofia. Senza dimenticare che digital libraries di testi, come la raccolta di classici della letteratura prodotti in seno al progetto Gutenberg, sono già esposti come linked data

[10]Si tratta del principale schema in uso in campo di markup di testi letterari e umanistici in senso ampio: http://www.tei-c.org.

[11]Projects using TEI: http://www.tei-c.org/Activities/Projects.

[12]Anche se non si può non notare che il neonato SAN (Sistema Archivistico Nazionale): http://san.beniculturali.it che vuole aggregare progetti digitali in ambito archivistico, riserva il concetto di digitalizzazione alla conversione di oggetti analogici, anche in termini di documenti di testo, nel solo formato immagine, riservando al metadato il solo ruolo descrittivo. La ragione evidentemente è che il numero di progetti di trascrizione annotata di documenti in campo archivistico è ancora limitata

[13]Progetto dell'Archivio di Stato di Prato: http://datini.archiviodistato.prato.it/margherita/index.htm.

[14]Progetto del Centro Scrineum dell'Università di Pavia: http://cdlm.unipv.it.

[15]Transcribe Bentham Transcription Desk: http://blogs.ucl.ac.uk/transcribe-bentham.

[16]http://www.discovery-project.eu/home.html.

sets[17] e già collegati ad altri data sets come DBpedia.[18]

Non è un caso poi se molte edizioni di testi si configurino come "archivi": Walt Whitman Archive,[19] Willa Cather Archive,[20] William Blake Archive,[21] Dante Gabriel Rossetti Archive,[22] Emily Dickinson's Archive;[23] si tratta di un processo che intende tradurre il concetto di edizione in quello di raccolta di documenti necessari alla classificazione del lavoro di un autore (Price). E l'edizione come archivio allarga il concetto di edizione a quello di base di conoscenza.

Un serbatoio quindi di informazione annotata che può essere arricchita e trasformata, diventare oggetto di riflessione alla ricerca di relazioni interne fra gli elementi e posta in collegamento con altre risorse per diventare una fonte di conoscenza. Se il processo di annotazione delle risorse a testo pieno, che ad oggi avviene nella maggior parte dei casi in forma manuale, potesse poi avvalersi di strumenti di riconoscimento automatico delle stringhe (information extraction, IE), e conseguente etichettatura, elaborati nel settore del natural language processing, come la named entity recognition, il sistema di costruzione di punti di accesso semantici ne trarrebbe ampio giovamento (per una visione d'insieme dei sistemi di IE si veda Chang et al.).

Ai vocabolari di annotazione in uso nel settore dell'edizione di testi, primo fra tutti lo schema Text Encoding Iniziative (TEI) basato

[17] Project Gutenberg Catalog: http://wifo5-03.informatik.uni-mannheim.de/gutendata.

[18] "DBpedia is a crowd-sourced community effort to extract structured information from Wikipedia and make this information available on the Web." http://dbpedia.org.

[19] http://www.whitmanarchive.org.

[20] http://cather.unl.edu.

[21] http://www.blakearchive.org/blake.

[22] http://www.rossettiarchive.org.

[23] http://www.emilydickinson.org.

sull'embedded markup XML, si aggiungono gli standard, vale a dire sets di metadati e relativi valori o ontologie, che identificano il sistema descrittivo delle risorse digitali in uso negli ambienti di gestione e trattamento del patrimonio culturale. A livello di metadati/ontologie, ovvero di element sets, il mondo degli archivi ha gli schemi EAD[24] e il già citato EAC-CPF, i musei hanno il CIDOC-CRM,[25] il Web, e i sistemi di esposizione di metadati, investono su DC[26] come strumento per la disseminazione. SKOS[27] è un modello in uso nel settore della costruzione di reti lessicali. FRBR[28] è un altro modello, standard dell'IFLA, che dalle biblioteche si sta estendendo ai diversi àmbiti della metadatazione di risorse in cui il processo di stratificazione, o il punto di vista multilivello, svolge un ruolo fondamentale nella descrizione dell'oggetto dell'analisi. E poi c'è Europeana che ha elaborato un data model finalizzato a raggruppare e mappare vari modelli concettuali e ontologie.[29] Al set di descrittori, elementi o classi, si aggiunge la questione dei valori. Altrettanto numerosi i vocabolari in uso nella forma della tassonomia o del thesaurus: p.e. AAT (Art and Architecture Thesaurus) del Getty, lo storico DDC (Dewey Decimal Classification), IconClass, GeoNames, Wordnet.[30] E poi esistono le authorities della Library of Congress[31] e il progetto

[24]Encoded Archival Description: http://www.loc.gov/ead.

[25]CIDOC - Conceptual Reference Model: http://www.cidoc-crm.org.

[26]Dublin Core: http://dublincore.org.

[27]Simplified Knowledge Organisation Systemhttp: http://www.w3.org/2004/02/skos.

[28]Functional Requirements for Bibliographic Records: http://www.ifla.org/publications/functional-requirements-for-bibliographic-records.

[29]Europeana Data Model (EDM) Documentation: http://pro.europeana.eu/edm-documentation.

[30]Un elenco completo dei value vocabularies si può leggere nel report del W3C Incubator Group del 25 ottobre 2011, *Library Linked Data Incubator Group: Datasets, Value Vocabularies, and Metadata Element Sets*: http://www.w3.org/2005/Incubator/lld/XGR-lld-vocabdataset-20111025.

[31]Library of Congress authority: http://authorities.loc.gov; Library of Congress

VIAF[32] che vogliono proporsi come descrittori univoci, anche in un'ottica linked data. E numerosi sono anche gli aggregatori di vocabolari, ontologie e linked data sets: dal Metadata Registry[33] al LOV,[34] da LOD Cloud[35] ai Semantic Web Search Engines[36] finalizzati al recupero di informazione semanticamente consistente. I principi del semantic web, e di linked data in particolare, si stanno imponendo come modello teorico e tecnologico di riferimento nel settore delle humanities e in particolare delle biblioteche, degli archivi e dei musei allo scopo di allargare le prospettive di interlinking fra risorse prodotte dagli istituti di conservazione (Guerrini e Possemato).[37]

Ovviamente l'esigenza nella rappresentazione di un dominio è usare standard condivisi sulla base delle regole condivise e rendere le descrizioni compatibili con altri domini e quindi altri standard. Grande lavoro sul cross-mapping e su problemi di allineamento si sta facendo (Haslhofer e Klas) e fin dal 1996 la molteplicità di standard di metadati è sentito come un problema (Day). Ma molte questioni sono ancora da risolvere.

Se dal punto di vista di metadati/ontologie e vocabolari il panorama è estremamente eterogeneo, dal punto di vista delle tecnologie, intese come linguaggi formali per la descrizione delle risorse, uno sforzo comune si sta invece registrando. XML, RDF, URI e OWL sono ormai termini comunemente in uso nel settore del digital cultural

Linked Data Service http://id.loc.gov.

[32]Virtual International Authority File: http://viaf.org.

[33]http://metadataregistry.org.

[34]Linked Open Vocabularies: http://lov.okfn.org/dataset/lov.

[35]Linking Open Data Cloud di Ckan: http://datahub.io/group/lodcloud.

[36]Un elenco si può consultare sul wiki del W3C sul semantic web, nell'àmbito delle attività della Task Force su linking open data: http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/SemanticWebSearchEngines.

[37]Come dimostra la bella raccolta di contributi del convegno *Global Interoperability and Linked Data in Libraries* tenutosi a Firenze il 18 e 19 giugno 2012 e i cui atti sono pubblicati da JLIS.it: http://leo.cilea.it/index.php/jlis/issue/view/536.

heritage. Che si produca un'annotazione embedded (p.e. XML/-TEI) o una annotazione stand-off ogni elemento interpretativo, che può diventare un authority record, deve essere identificato univocamente. Le tecnologie del semantic web aiutano a far fronte al problema dell'identificazione univoca e della sua modalità di espressione attraverso il meccanismo degli URIs. A livello URI è possibile attribuire ad ogni entità una serie di informazioni, mettendo in relazione tale entità con altri URIs attraverso asserzioni RDF, che possono anche prevedere l'utilizzo di predicati ontologici esistenti. Sempre attraverso le stesso meccanismo se si dispone degli URIs di altre risorse Web, magari esposti come data sets, è possibile creare relazioni fra gli elementi annotati e le altre risorse che condividono con le prime determinate features. Tale annotazione, che può venire quindi trasformata in questo modo in data set, può essere esposta su Web attraverso grafi RDF e di conseguenza essere visibile ad altri utenti. Anche il testo dell'edizione, esposto come grafo RDF, può essere mostrato, e volendo anche popolato, da altri ricercatori.[38] In questo contesto un ruolo importante ricopre il framework OAC (Open Annotation Collaboration)[39] come strategia per la gestione delle relazioni fra documento e annotazione e per l'interoperabilità fra annotazioni in prospettiva RDF (Barbera et al.). Il processo di estrazione di triple RDF da file, che utilizzano per esempio il vocabolario TEI, attraverso il modello OAC risulta peraltro un ambito di riflessione critica estremamente interessante nell'ambito delle digital humanities (Jordanous, Stanley e Tupman).

[38]Sul ruolo delle tecnologie nell'ambito linked data si vedano guide e tutoriali sul sito: http://linkeddata.org.

[39]Si veda il recente Open Annotation Data Model: http://www.openannotation.org/spec/core.

3 Le fasi del processo

La costruzione di authority files come raccolta di dati controllati che vengono estratti delle edizioni di documenti si scontra con l'importanza del contesto in cui ogni authority compare e quindi con il dominio di riferimento in cui quell'authority può essere ricompreso. Il problema si articola su tre livelli: come descrivere gli elementi dell'annotazione, che possono essere le entries di un authority record; come creare le relazioni fra tali elementi, che può diventare un sistema di approfondimento del concetto di authority come raccolta di dati contestuali; come far dialogare tali elementi e quindi l'edizione, con il WWW attraverso linked data. E quindi come trasformare autority files, che nascono da un contesto testuale e sono relativi ad un dominio, in linked data sets autoesplicativi, coerenti e appropriati e in grado quindi di dialogare con altre risorse correlate. L'informazione che proviene dai testi delle edizioni può fornire importanti concetti che possono essere formalizzati per la costruzione di basi di conoscenza.

Partiamo da un caso di studio per esemplificare il procedimento: un'edizione digitale di una raccolta di lettere manoscritte, conservate in istituzioni archivistiche e in biblioteche nazionali, ricevute e inviate, nel corso del XV secolo, dal/al copista e libraio fiorentino Vespasiano da Bisticci (Tomasi, «L'edizione digitale e la rappresentazione della conoscenza. Un esempio: Vespasiano da Bisticci e le sue lettere»; «Digital editions between embedded markup and external representation. A case study: Vespasiano da Bisticci's Letters»).

3.1 Elementi e valori

Il primo problema in un approccio finalizzato a stabilire descrittori e relativi valori per la creazione di authority files è la selezione dei metadati quindi la scelta di un vocabolario controllato sia a livello

di elementi che di valori. Due sono quindi i piani su cui ragionare: quali element sets è più opportuno utilizzare per esprimere il punto di vista dell'interprete sulla fonte, che rappresenta il modello; quali value vocabularies sono i più appropriati per esprimere il valore di un elemento. Supponiamo di voler esprimere il seguente concetto, o la seguente asserzione, come lo possiamo dedurre da una lettera di Vespasiano da Bisticci a Piero de' Medici:[40]

> Vespasiano da Bisticci ha copiato le Vite di Plutarco per Piero de' Medici a Firenze nel 1441

Possiamo formalizzare il concetto iniziando a scomporne gli elementi costitutivi secondo il modello "who, where, when, what" e impiegando i nomi di elementi, o le denominazioni delle classi, come stabiliti dai più comuni modelli di metadati o ontologie (per esempio i già citati TEI, CIDOC-CRM, DC, EAC-CPF, EDM).[41]

In un approccio finalizzato a ridefinire il ruolo e la funzione di un authority come stringa estratta da un contesto specifico d'uso e relativa ad un altrettanto specifico dominio i problemi riguardano sia la definizione dei nomi delle etichette descrittive che i valori associati.

Da questo esempio è facile dedurre che denominazioni di elementi diversi esprimono in realtà lo stesso concetto (es. "placename"

[40] Supponendo di voler tradurre la forma attestata di nomi di persona, date, luoghi ed eventi in un documento nella corrispettiva forma controllata come stabilita da una authority condivisa.

[41] Senza ambire ovviamente ad una mappatura dei modelli o all'esaustività della rappresentazione. Alcuni valori potrebbero essere suscettibili di ulteriori scomposizioni (e.g. manuscript-of-Plutarchus-Vitae). Peraltro TEI sta lavorando al mapping, come si può leggere sul wiki dedicato all'attività dello Special Interest Group (SIG) sulle ontologie: http://wiki.tei-c.org/index.php/SIG:Ontologies, in particolare TEI su CIDOC-CRM (Eide e Ore). Grande lavoro sul mapping ha poi fatto Europeana per il suo data model, fornendo peraltro linee guida specifiche. Le *Mapping Guidelines* v1.0.1 (del 24.02.2012) si possono consultare all'indirizzo: http://pro.europeana.eu/documents/900548/ea68f42d-32f6-4900-91e9-ef18006d652e.

Elemento/Classe	**Valore**
persname/creator/actor/agent/person	Bisticci_Vespasiano_da
persname/person	Medici_Piero_de
placename/place_appellation/place	Firenze – Florence
date	1441
event	copy-of-codex
object/physical_thing	manuscript-of-Plutarchus-Vitae

e "place_appellation") e che i valori associati non sempre sono formulabili secondo i precetti di un vocabolario controllato (es. un evento). Diciamo che a livello di mapping molte ambiguità terminologiche sono risolvibili, anche se non sempre lo stesso elemento è interpretato esattamente con lo stesso significato dai modelli in uso (e questo deriva principalmente dalle circostanze di implementazione del modello e dal contesto d'impiego, e.g. "actor" in TEI è utilizzato in modo diverso rispetto al CIDOC-CRM).

Per quanto riguarda i valori esistono, come noto, numerosi vocabolari controllati (già menzionati in precedenza: per le persone, ma anche per i titoli e le keywords, ci sono per esempio le authorities della Library of Congress, per i luoghi il database GeoNames, per i soggetti in Italia c'è il nuovo soggettario della BNCF, ma esiste anche Wordnet a livello internazionale, per le date lo standard ISO 8601). Ma non è detto che tali vocabolari siano sufficienti ad esprimere ogni valore associato all'elemento oltre a soddisfare le esigenze di comunità diverse (sul vocabulary alignment, in modo particolare per i soggetti, si veda, storicamente, Doerr).

Come in un sistema di authority attenzione speciale la si qui vuole dedicare al concetto di persona. Si tratta di un elemento su cui numerosi modelli di metadazione hanno riflettuto. Il primo problema nella definizione di un authority file per le persone è la definizione della forma accettata del nome. E su questo problema una volta che ogni progetto dichiari a quale istituzione deputata

a stabilire il controllo d'autorità si rivolge (es. l'Istituto Centrale per il Catalogo Unico per l'Italia o VIAF a livello internazionale) è possibile rivolvere la questione, anche se ci dovrebbe essere condivisione a livello internazionale circa chi debba ricoprire questo ruolo di garante del controllo di autorità. A cui possiamo aggiungere che le forme attestate nei documenti possono fornire utili forme varianti che possono arricchire authorities esistenti.[42]

Ma particolarmente importante è la connotazione del concetto "persona" nei diversi modelli di metadazione. Diciamo che è evidente che un'etichetta come "EDM:agent" o "CIDOC-CRM_E39:actor" determina un'azione della persona ed è quindi altro rispetto a "person". Allo stesso modo "DC:creator" determina una funzione o meglio un ruolo. Particolare attenzione andrà allora prestata alla descrizione del concetto di persona in quanto il ruolo, la funzione e l'azione sono caratteristiche che possono cambiare a seconda del contesto testuale in cui l'entità compare. Ecco quindi che, astraendo, l'authority, come stringa estratta da un concetto espresso dal documento, inizia a configurarsi: la persona identificata ha un ruolo e ha svolto una specifica funzione che ha portato alla realizzazione di qualcosa a favore di un'altra persona in un certo luogo e in un certa data come attestato dalla fonte in cui l'entità compare.

3.2 Relazioni fra elementi o classi

Passiamo quindi dalla riflessione in termini di accoppiate elemento-valore a quella di asserzione in termini soggetto/predicato/oggetto,

[42] Per esempio VIAF attesta diverse forme di Vespasiano da Bisticci (il cui VIAF ID è 76466245 e il permalink http://viaf.org/viaf/76466245): Vespasiano, da Bisticci, 1421-1498; Vespasiano da Bisticci, Fiorentino, 1421-1498; Vespasiano, da Bisticci, ca. 1421-1498; Vespasiano Da Bisticci, Fiorentino; Bisticci, Vespasiano Da. Dalla collezione di lettere in questione desumiamo invece che Vespasiano si firma sempre come "Vespasiano di Filippo".

secondo i precetti di RDF. Ovviamente nel momento in cui si ontologizza la conoscenza alcune classi diventano proprietà e i valori, intesi come risorse, diventano istanze potenzialmente dotate di URIs e quindi univocamente identificabili.

Le relazioni, o meglio la definizione delle proprietà, diventa un modo per esplicitare formalmente le interpretazioni dell'editore critico. La lettura del testo da parte dell'editore comporta quindi la determinazione del sistema di collegamenti. Il contesto in cui una persona, un luogo o una data sono inseriti fa di quell'istanza una fonte di informazioni proprio in quanto contestualizzata rispetto a quella specifica situazione testuale. La stessa istanza potrebbe assumere un valore diverso quando calata in un differente contesto.

A questo problema si aggiunge la modalità della dichiarazione delle proprietà, vale a dire la definizione dei criteri con cui esprimere le relazioni fra gli elementi annotati, ovvero la scelta dei predicati ontologici e la verifica degli esistenti, allo scopo di comprendere se altre ontologie soddisfino i bisogni interpretativi dell'editore critico. Prendiamo un caso semplice. La relazione fra una persona, identificata da un elemento "persname", e associata ad un letterale in vocabolario controllato, e il luogo in cui quella persona è nata, utilizzando gli elementi TEI e la proprietà "birth":

TEI:persname#Bisticci_Vespasiano_da
birth
TEI:placename#Florence

Caso già più particolare potrebbe essere il seguente concetto:

TEI:persname#Bisticci_Vespasiano_da
copyied-where
TEI:placename#Florence

In questo ultimo caso si sta esprimendo una proprietà che collega le stesse istanze precedenti (Vespasiano e Firenze), ad identificare la

relazione fra una persona e un luogo come desunta da uno specifico contesto testuale in cui la proprietà (luogo in cui è avvenuta la copia di un codice) è specifica per l'occorrenza che si vuole documentare. Ma potremmo anche dire (in una linearizzazione non standardizzata):

> actor/person#Bisticci_Vespasiano_da
> copyied-for
> addressee/person#Medici_Piero_de

A specificare anche i ruoli ("actor" e "addressee") che diverse persone hanno in uno specifico contesto in cui accade un determinato evento (una copia effettuata da un individuo per un altro individuo) in un dato momento.

Ovviamente il problema della compatibilità e dell'interscambio fra i modelli concettuali se deve avvenire in termini di classi e sottoclassi deve avvenire anche a livello di predicati. Sarà dunque necessario mappare i predicati utilizzati in uno specifico contesto con i predicati affini utilizzati in altri modelli affinché la collezione sia davvero interoperabile a livello semantico. Il data model proposto da Europeana, il già citato EDM (Doerr et al.), può essere un riferimento, anche perché per sua stessa natura deve confrontarsi con standard di metadati diversi e renderli compatibili attraverso la definizione di uno schema unico condiviso (Peroni, Tomasi e Vitali).

Per quanto riguarda le relazioni fra persone ISAAR-CPF è un buon modello di riferimento. In ISAAR-CPF il concetto di relazione lega fra loro i soggetti produttori (in senso estensivo le persone) ma anche i soggetti produttori con le risorse prodotte. Ogni relazione fra soggetti può essere classificata (es. gerarchica, cronologica, familiare, associativa), descritta (volendo utilizzando anche un vocabolario controllato) e datata (impiegando p.e. una convenzione come ISO 8601). Allo stesso modo le relazioni fra un soggetto e una risorsa possono essere tipizzate, può essere descritta la natura della relazione e

fornita una datazione. EAC-CPF acquisisce le specifiche ISAAR-CPF e propone un "eac:relations" che si basa sul principio degli "agents" come soggetti produttori e dei collegamenti fra soggetti intesi come unità complesse ("entities"), fornendo poi gli strumenti per specificare la funzione della relazione ("functionRelation"), e per determinare e rappresentare relazioni fra soggetti e risorse correlate ("resourceRelation").

Per ragionare in termini di authority records oltre ad EAC-CPF dovrebbero essere seguite le indicazioni di MADS[43] che, nel definire un modello di authority record, insiste sul problema delle relazioni fra persone e RDA[44] che, fra le altre cose, e sulla scorta di FRBR, ragiona sul concetto di persona, sia a livello di attributes, che di relationships.[45] L'authority estratta da un documento diventa quindi un'entità più strutturata che prevede, oltre a forme controllate delle entries, anche la serie delle relazioni necessarie a documentare un contesto. Si inizia così a semantizzare con collegamenti tipizzati che determinano una nuova authority come punto di accesso ai concetti intesi come relazioni fra istanze contestuali, in cui la fonte svolge un ruolo fondamentale nella definizione del concetto.

3.2.1 Relazioni con linked data sets

Affinché authority records così configurati possano essere interoperabili anche a livello semantico è necessario porli in dialogo con la realtà del WWW. Questo significa trasformare le authorities in data sets e rendere questi ultimi pubblicamente disponibili; ma significa anche conoscere ed utilizzare data sets esistenti qualora ci siano

[43] Metadata Authority Description Schema: http://www.loc.gov/standards/mads.

[44] Resource Description & Access: http: http://www.rda-jsc.org/rda.html.

[45] Un bel progetto denominato SNAC (Larson e Janakiraman) è un esempio prototipale di riflessione sul concetto di persona e sulle associazioni: http://socialarchive.iath.virginia.edu. L'accesso al prototipo all'indirizzo: http://socialarchive.iath.virginia.edu/xtf/search.

possibili collegamenti, per aprire il concetto di relazione a quello di contesto esteso, determinato dal collegamento. Ovviamente con RDF e URIs dereferenziabili creare un data sets non è operazione complessa. E la scelta degli URIs può essere fatta consapevolmente impiegando data sets già esistenti e certificati (es. i già citati VIAF per le forme controllate dei nomi, il progetto Gutenberg per autori e testi, LC Linked Data Service per gli authority records, o ancora DBpedia per i nomi e Wordnet per le keywords).[46] Più complesso concettualmente riconoscere che il data set documenta occorrenze relative ad uno specifico dominio e relative ad un determinato contesto testuale in cui un'entità occorre. La complessità deriva dal fatto che se la proprietà "owl:same-as", utilizzata comunemente per definire forme di corrispondenza fra entità, aiuta a documentare l'esistenza di URIs affini, bisogna ricordare che la stessa entità, se calata in un diverso contesto testuale, potrebbe veicolare un diverso concetto.

Certamente non bisogna dimenticare che la vera interoperabilità è determinata dall'impiego di risorse già formalizzate e che la moltiplicazione di URIs relativi alla stessa istanza inficia il processo di dialogo. Quindi certamente creare collegamenti fra una risorsa e la sua forma standardizzata, o acquisirne l'URI (authority control via permalink), è importante, anche se è necessario sia esito di un ragionamento che tiene conto della specificità in cui la risorsa è calata. Ne deriva che il data set prodotto da ogni edizione produce documentazione relativa a istanze contestuali e che quindi la relazione fra data sets è determinata dalla condivisione di un concetto non di

[46]Un elenco completo dei data sets ad oggi disponibili, e dei relativi URL di progetto e URIs di risorse, si può leggere nella sezione della Task Force del W3C SWEO Community Project: *Linking Open Data on the Semantic Web*, http://www.w3.org/wiki/TaskForces/CommunityProjects/LinkingOpenData/DataSets oppure sul già citato Ckan, "a registry of open data and content packages provided by the Open Knowledge Foundation": http://datahub.io.

semplici stringhe.

Certamente tanto più aumenteranno i data sets esistenti e verranno rese disponibili le triple su WWW, aprendo le risorse al dialogo e non mantenendole "siloed", tanto più la rete della conoscenza diventerà efficace. Non bisogna poi dimenticare che se linked data è una modalità di rappresentazione dell'informazione che ambisce alla costruzione di relazioni la comunità del semantic web, e dell'intelligenza artificiale in particolare, coglie ancora dei limiti derivati dall'assenza di una "upper level ontology" che davvero agevoli forme di ragionamento automatico (Jain et al.).

4 Conclusioni

Scopo del presente lavoro è quindi di aprire una strada verso l'edizione digitale come raccolta testuale da cui acquisire dati che possano essere rappresentati come un nuovo modello di authority record, in cui cioè le stringhe annotate ed estratte dai testi pieni delle edizioni diventino punti di accesso al bagaglio informativo trasmesso dai documenti e in cui la fonte dove l'entità appare è determinante a stabilire il significato. In prima battuta le informazioni già etichettate possono essere estratte da testi marcati, che già presentano un primo livello di descrizione e forniscono le entries. Queste ultime diventano un'authority, arricchita con altre entità correlate a diversi livelli, e la relazione rappresenta una nuova authority. Esporre questi dati sotto forma di open data sets garantisce una ricchezza di risorse aggiuntive per l'interscambio; utilizzare data sets certificati per costruire relazioni e collegamenti deve fare i conti con le diverse situazioni in cui le entità occorrono. Il principio del contesto testuale in questa argomentazione, anche secondo le modalità con cui tale espressione viene utilizzata in campo archivistico, è fondamentale per la costruzione di nuove authorities, che documentano il domi-

nio in cui le entità occorrono. E l'interscambio è determinato dalla condivisione di concetti. Il concetto diventa un nuovo strumento per esplorare i contenuti espressi dai documenti, trasformando le authorities in punti di accesso semantici. Questo processo, oltre a valorizzare i documenti digitali, fornisce nuove fonti utili per l'arricchimento di liste di autorità e fornisce una nuova metodologia di esplorazione del full-text dei documenti; l'authority si viene a configurare come un record complesso in cui contesto e dominio determinano nuovi concetti.

Riferimenti bibliografici

Aloia, Nicola, Cesare Concordia e Carlo Meghini. «Europeana v1.0». *Digital Libraries and Archives*. A cura di Maristella Agosti, et al. Vol. 249. Communications in Computer and Information Science. Berlin, Heidelberg: Springer, 2011. 127–129. <http://link.springer.com/content/pdf/10.1007\%2F978-3-642-27302-5_16.pdf#page-1>. (Cit. a p. 25).

Barbera, Michele, et al. «Annotating Digital Libraries and Electronic Editions in a Collaborative and Semantic Perspective». *Digital Libraries and Archives*. Berlin, Heidelberg: Springer, 2012. 45–56. <http://link.springer.com/chapter/10.1007/978-3-642-35834-0_7>. (Cit. a p. 31).

Chang, Chia-Hui, et al. «A Survey of Web Information Extraction Systems». *IEEE Transactions on Knowledge and Data Engineering* 18.10. (2006): 1411–1428. (Cit. a p. 28).

Clement, Tanya. «Knowledge Representation and Digital Scholarly Editions in Theory and Practice». *Journal of the Text Encoding Initiative* 1. DOI: 10.4000/jtei.203. (2011). (Cit. a p. 25).

Coombs, James H., Allen H. Renear e Steven J. DeRose. «Markup systems and the future of scholarly text processing». *Communications of the ACM* 30.11. DOI: 10.1145/32206.32209. (1987): 933–947. (Cit. a p. 21).

Crupi, Gianfranco. «Beyond the Pillars of Hercules: Linked data and cultural heritage». *JLIS.it* 4.1. DOI: 10.4403/jlis.it-8587. (2013). (Cit. a p. 25).

Day, Michael. *Mapping between metadata formats*. 1996. <http://www.ukoln.ac.uk/metadata/interoperability>. (Cit. a p. 30).

D'Iorio, Paolo e Michele Barbera. «Scholarsource: A Digital Infrastructure for the Humanities». *Switching Codes. Thinking through New Technology in the Humanities and the Arts*. A cura di Roderick Coover e Thomas Bartscherer. Chicago: University of Chicago Press, 2011. 61–87. (Cit. a p. 27).

Doerr, Martin. «Semantic Problems of Thesaurus Mapping». *Journal of Digital Information* 1.8. (2001). <http://journals.tdl.org/jodi/index.php/jodi/article/viewArticle/31/32>.

Doerr, Martin, et al. «The Europeana Data Model (EDM)». *Proceedings of 76th IFLA General conference and Assembly*. 2010. <http://conference.ifla.org/past/ifla76/149-doerr-en.pdf>. (Cit. a p. 37).

Eide, Øyvind e Christian-Emil Ore. «TEI and cultural heritage ontologies: Exchange of information?» *Literary and Linguist Computing* 24.2. (2009): 161–172. (Cit. a p. 33).

Guerrini, Mauro. «Global Interoperability and Linked Data in Libraries: Special issue.» *JLIS.it* 4.1. (2013). <http://leo.cilea.it/index.php/jlis/issue/view/536>.

Guerrini, Mauro e Tiziana Possemato. «Linked data: a new alphabet for the semantic web». *JLIS.it* 4.1. DOI: 10.4403/jlis.it-6305. (2013). (Cit. a p. 30).

Haslhofer, Bernhard e Antoine Isaac. «data.europeana.eu - The Europeana Linked Open Data Pilot». *DCMI International Conference on Dublin Core and Metadata Applications*. The Hague, The Netherlands. 2011.

Haslhofer, Bernhard e Wolfgang Klas. «A survey of techniques for achieving metadata interoperability». *ACM Computing Surveys* 42.2. DOI: 10.1145/1667062.1667064. (2010): 7:1–7:37. (Cit. a p. 30).

Jain, Prateek, et al. «Linked Data Is Merely More Data». *Linked Data Meets Artificial Intelligence*. 2010. <http://www.aaai.org/ocs/index.php/SSS/SSS10/paper/view/1130>. (Cit. a p. 40).

Jordanous, Anna, Alan Stanley e Charlotte Tupman. «Contemporary transformation of ancient documents for recording and retrieving maximum information: when one form of markup is not enough». *Balisage: The Markup Conference*. 2012. (Cit. a p. 31).

Larson, Ray R. e Krishna Janakiraman. «Connecting Archival Collections: The Social Networks and Archival Context Project». *Research and Advanced Technology for Digital Libraries*. A cura di Stefan Gradmann, et al. Berlin, Heidelberg: Springer, 2011. 3–14. <http://link.springer.com/chapter/10.1007/978-3-642-24469-8_3>. (Cit. a p. 38).

Lee, Christopher A. «A framework for contextual information in digital collections». *Journal of Documentation* 67.1. DOI: 10.1108/00220411111105470. (2011): 95–143. (Cit. a p. 24).

Mazzini, Silvia e Francesca Ricci. «EAC-CPF Ontology and Linked Archival Data». *Proceedings of the 1st International Workshop on Semantic Digital Archives*. Berlin: CEUR, 2011. (Cit. a p. 24).

Peroni, Silvio, Francesca Tomasi e Fabio Vitali. «Reflecting on the Europeana Data Model». *Digital Libraries and Archives*. A cura di Maristella Agosti, et al. Communications in Computer and Information Science 354. Berlin, Heidelberg: Springer, 2012. 228–240. <http://link.springer.com/chapter/10.1007/978-3-642-35834-0_23>. (Cit. a p. 37).

Pitti, Daniel. «Creator Description: Encoded Archival Context». *Authority control in organizing and accessing information: definition and international experience*. A cura di Arlene G. Taylor, et al. Binghamton N.Y.: Haworth Information Press, 2004. 201–226. (Cit. a p. 24).

Price, Kenneth M. «Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?» *DHQ* 3.3. (2009). <http://www.digitalhumanities.org/dhq/vol/3/3/000053/000053.html>. (Cit. a p. 28).

Sahle, Patrick. *A catalog of Digital Scholarly Editions*. 2013. (Cit. a p. 26).

Siemens, Ray, et al. «Toward modeling the social edition: An approach to understanding the electronic scholarly edition in the context of new and emerging social media». 27.4. DOI: 10.1093/llc/fqs013. (2012): 445–461. (Cit. a p. 27).

Tomasi, Francesca. «Digital editions between embedded markup and external representation. A case study: Vespasiano da Bisticci's Letters». *Quaderni digilab* 2.1. (2012): 201 –218. <http://digilab-epub.uniroma1.it/index.php/Quaderni_DigiLab/article/view/24>. (Cit. a p. 32).

——. «L'edizione digitale e la rappresentazione della conoscenza. Un esempio: Vespasiano da Bisticci e le sue lettere». *Ecdotica* 9. (2012). in print. (Cit. a p. 32).

FRANCESCA TOMASI, Università di Bologna.
francesca.tomasi@unibo.it

Tomasi, F. "Le edizioni digitali come nuovo modello per dati di autorità concettuali". *JLIS.it*. Vol. 4, n. 2 (Luglio/July 2013): Art: #8808. DOI: 10.4403/jlis.it-8808. Web.

ABSTRACT: Projects related to cultural heritage enhancement are facing a gradual transition from the description of the sources, at the level of metadata, to their digitization. When this heritage is textual a special attention is recognized to digitization as annotated or "marked-up" transcription, having the aim of textual or documentary edition. Each feature of a document that can be element of annotation - and is therefore subject of interpretation - takes the form of an authority data to be analyzed under the different aspects that attest the specific instance of the element in context. Tools of description of resources, as product of context and domain, contribute to transform the edition of a document in a knowledge base. Semantic Web and Linked Data provides the theoretical and technological tools to convert siloed authority files, which represent the conceptual or semantic access points to digital editions, in interoperable resources.

KEYWORDS: Authority control; Digital editions; Linked data; Semantic indexing; Semantic web.

Submission: 2013-03-02
Accettazione: 2013-05-16
Pubblicazione: 2013-07-01